Information Processing and Management 59 (2022) 102799

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# Fine-grained citation count prediction via a transformer-based model with among-attention mechanism

Shengzhi Huang [a,b,1], Yong Huang [a,b,1], Yi Bu [c], Wei Lu [a,b], Jiajia Qian [a,b], Dan Wang [a,b,*]

[a] School of Information Management, Wuhan University, Wuhan, Hubei, China
[b] Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan, Hubei, China
[c] Department of Information Management, Peking University, Beijing, China

ARTICLE INFO

*Keywords:*
Citation count prediction
Functional structure
Neural network
Content-based citation analysis

ABSTRACT

Previous studies have confirmed that citation mention and location reveal different contributions of the cited articles, and that both are significant in scientific research evaluation. However, traditional citation count prediction only focuses on predicting citation frequency. In this paper, we propose a novel fine-grained citation count prediction task (FGCCP), which aims to predict in-text citation count from each structural function of a paper separately. Specifically, we treated this task as a "sequence to sequence" issue and a multi-task learning job, in which both the inputs and the outputs are based on the sequence pattern of citations from different structural functions. To fulfill FGCCP, we proposed a transformer-based model (i.e. MTAT) in which a novel among-attention mechanism is employed. Based on an empirical study of full-text documents from PubMed Central Open Access Subset, our model achieves satisfactory prediction accuracy, and surpasses common machine learning and deep learning models on FGCCP. Moreover, we also discuss the potential role of the among-attention mechanism and the reason why our proposed model outperforms state-of-the-art strategies. FGCCP may provide more detailed decision-making evidence and evaluation basis for researchers in scientific research evaluation. In addition, MTAT is a general model which can be easily deployed in other multi-task learning jobs.

## 1. Introduction

Citation count is the simplest yet most widely used indictor to approximate the scientific impact of publications, authors, and institutions (Bu, Lu, Wu, Chen, & Huang, 2021; Cao, Chen, & Liu, 2016; Lu, Ding, & Zhang, 2017; Stegehuis, Litvak, & Waltman, 2015, p.; Yu, Yu, Li, & Wang, 2014). Many decisions with regard to the allocation of funding and the promotions of researchers are closely related to this indicator (Mazloumian, 2012; Mingers & Burrell, 2006). However, the accumulation of citations consumes time and leads to a delay for research evaluation (Akella, Alhoori, Kondamudi, Freeman, & Zhou, 2021; Ruan, Zhu, Li, & Cheng, 2020). Science funding providers and research evaluators are eager to know which papers may generate high impact in the future, and one of the most intuitive ways to discover this is to predict the citation count of a given paper shortly after its publication.

* Corresponding author at: School of Information Management, Wuhan University, Wuhan, Hubei, China.
*E-mail addresses:* ShengzhiHuang@whu.edu.cn (S. Huang), yonghuang1991@whu.edu.cn (Y. Huang), buyi@pku.edu.cn (Y. Bu), weilu@whu.edu.cn (W. Lu), jiajiaqian@whu.edu.cn (J. Qian), info_wd@126.com (D. Wang).
[1] Shengzhi Huang and Yong Huang contributed equally to this work.

https://doi.org/10.1016/j.ipm.2021.102799
Received 25 July 2021; Received in revised form 14 October 2021; Accepted 17 October 2021
Available online 9 November 2021
0306-4573/

Citation count prediction was an important direction under traditional citation analysis, which treats all citations equally. However, some studies confirmed that citations from different locations of a publication may play different roles and show distinct academic contributions of the cited articles (Elkiss et al., 2008; Lu et al., 2017; Suppe, 1998). Zhao and Strotmann (2020), for instance, pointed out that citations in Method, Results, Discussion, and Conclusions sections are often quite important to the citing article compared with citations in other sections of a publication. Moreover, methodology-oriented papers tend to be more frequently cited in the Method section of a citing paper (Hu, Chen, & Liu, 2013), and are likely to be more highly cited than other types of articles (Boyack, van Eck, Colavizza, & Waltman, 2018). In addition, the number of in-text citations, to some extent, is more representative regarding the contribution of a reference to the citing paper (Herlach, 1978; Pak, Wang, & Yu, 2020; Voos & Dagaev, 1976; Zhao & Strotmann, 2020). Therefore, we maintain that a fine-grained citation count prediction (FGCCP), which predicts citation count and location simultaneously, is of great significance to scientific research evaluation. In this paper, we followed the definition of structural functions and functional structure proposed by Lu, Huang, Bu, and Cheng (2018), by which citation locations of different papers can be determined in a uniformed and meaningful way. The research objective of this study is to propose a new method to predict in-text citation count from different structural functions of a paper, which may provide more abundant citation information for decision-makers.

More specifically, we treated FGCCP as a “sequence to sequence” issue, in which both the inputs and the outputs are based on the sequence pattern of citations. Given the in-text citation count of a paper in recent $S$ years, the goal of FGCCP is to predict in-text citation count from distinct structural functions of the paper in the next $S'$ years. Therefore, it is fundamentally a regression task and can also be seen as a multi-task learning job. To fulfill FGCCP, we also proposed a transformer-based model (i.e. MTAT), where the transformer (Vaswani et al., 2017) is employed, and a novel among-attention mechanism is utilized to connect multi transformers closely. Our experimental results on full-text documents collected from PubMed Central Open Access Subset confirmed that our model achieved satisfactory prediction performance on FGCCP. Moreover, MTAT has been compared with common methods in citation count prediction, which shows that our model is the most effective in FGCCP.

This research has the following theoretical and practical implications. First, compared with traditional citation count prediction, FGCCP is a more challenging task, which aims to predict in-text citation count and citation location simultaneously. Hence, it predicts the scholarly impact of a paper in more detail, and may provide more evidence for researchers to distinguish the impact of publications. Second, our proposed model achieved the satisfactory prediction accuracy, and surpassed common machine learning and deep learning algorithms on FGCCP. The among-attention mechanism utilized in MTAT has been confirmed to be effective, and may grasp the internal relationship among multi-tasks, so as to improve the overall performance on FGCCP. In addition, MTAT is a general model, which can also be used in other multi-task learning jobs.

The rest of this article is arranged as follows. First, we discuss related works about content-based citation analysis and citation count prediction. Then, we describe the prediction model and the dataset used in this article, which is followed by experimental settings and prediction results. Finally, we present theoretical and practical implications, limitations, and points for future research.

## 2. Background

### 2.1. Content-based citation analysis

Compared with traditional citation analysis, which treats all citations equally, content-based citation analysis focuses on analyzing the features of citation contexts (e.g. citation location and citation mention) to provide insight into the differences among citations (Ding et al., 2013, 2014; Lu et al., 2017).

Scientific articles are structured to clearly convey their topics, and each section of an article has its own special communicative function (Lu et al., 2018; Thelwall, 2019; Zhang, 2012). Hence, many studies have investigated citation location in citation analysis. Voos and Dagaev (1976) first addressed the issue of treating all citations equally. After examining citation location and citation mention for four highly cited articles, they concluded that the number of times a reference is mentioned and the citation location of that reference can reflect the contribution of the cited paper to a citing paper. Herlach (1978) also considered that a reference cited in different sections of the citing paper makes a greater contribution than references mentioned only once in the citing paper. Ding, Liu, Guo, and Cronin (2013) investigated the distribution of citation location in 866 information science articles, and found that highly cited papers appeared most often in the Introduction and Background sections of a citing paper. Hu et al. (2013) visualized citation distributions in 350 journal papers published in the Journal of Informetrics, and found that the first section of a paper possesses the most citations. Lu et al. (2017) employed content-based citation analysis to analyze the dynamics of one highly cited article, Hirsch’s “h-index” article. After checking the distribution of citation location of that article from 2006 to 2014, they identified three stages of the paper, and its impact continued to shift over the nine years. In this study, the definition of structural functions and functional structure (Lu et al., 2018) are employed to determine citation locations of cited papers in a uniformed and meaningful way. Functional structure consists of a variety of structural functions, and each structural function is a group of section headers with the similar communicative function. Citations from different structural functions of a citing paper may have different importance (Wan & Liu, 2014).

Many studies show that the number of in-text citations (i.e. citation mention) is significant in evaluating the scientific contribution of a reference to the citing paper (Ding et al., 2013; Herlach, 1978; Pak et al., 2020; Voos & Dagaev, 1976; Zhao & Strotmann, 2020). Boyack et al. (2018) examined characteristics of in-text citations in over five million publications collected from the PubMed Central Open Access Subset and Elsevier journals, and found that references mentioned only once tend to receive more citations than those mentioned multiple times. Pak et al. (2020) presented two counting methods (i.e. full counting and fractional counting) to analyze in-text citations. Their results show that most in-text citations are mentioned alone in a citation sentence, and most references having no independent mentions are mentioned only once in the citing paper. In addition, Zhao and Strotmann (2020) proposed a citation counting method which filters out in-text citations from the Introduction and Background sections before weighting the remaining citations. They tested the method on full-text documents from PubMed Central, and found their counting method makes essential citations stand out more. In this study, we employed the full counting method (Ding et al., 2013; Pak et al., 2020) to count citation frequency from different structural functions, separately.

### 2.2. Citation count prediction

Many researchers have focused on measuring and predicting the scholarly impact of publications, and citation count prediction is one of the most intuitive way to achieve this goal. Stochastic models, differential equation models, machine learning models, and deep learning models have been widely utilized in citation count prediction and have achieved fruitful results.

For a long time, stochastic methods have been used in bibliometrics and scientometrics (Glänzel & Schubert, 1995), such as in citation count prediction. Burrell (2002, 2003) employed a mixture of nonhomogeneous Poisson processes to predict citation behavior. In a mathematically precise way, he confirmed that, the longer an article has been uncited, the less likely it will be cited in the future. Mingers and Burrell (2006) regarded the citation process as a gamma mixture of Poisson processes. Under their assumptions, the distribution of the number of citations of a paper follows the negative binomial distribution, and their model achieved satisfactory performance on over 600 papers from six management science journals. Based on the Matthew effect, inherent quality of papers, and citation life cycle, Wang, Song, and Barabási (2013) proposed a differential equation model (WSB) for fitting the citation dynamics of individual papers. Their experimental results confirmed that the WSB can fit the citation trajectories of articles from different journals and disciplines. Based on the Hawk process, Bai, Zhang, and Lee (2019) proposed a paper potential index (PPI) model which involves four factors (i.e. inherent quality of papers, citation life cycle, early citations, and early citers' impact). Their PPI model achieves excellent predictive performance, and better explains the citation process. Although the above citation models based on point process or differential equation help researchers gain insight into the citation process, these models generally are constrained by priori hypotheses, which limits their generality.

Machine learning algorithms are commonly utilized in citation count prediction (Lu et al., 2021; Ruan et al., 2020) and have also yielded substantial results. For instance, Yu et al. (2014) analyzed paper features, journal features, author features, and citation features of papers, and utilized stepwise multiple regression to select appropriate features for predicting citation frequency. Onodera and Yoshikane (2015) also examined various extrinsic factors of an article, and employed negative binomial multiple regression to predict citation count. They confirmed that the proportion of the references within three and five years, and the number of references is especially important features. In a similar vein, Djokoto et al. (2020) utilized a negative binomial regression model and Poisson regression model to identify the drivers of citations of frontier application publications on Ghana. Stegehuis et al. (2015) employed journal impact and early citations as predictors, and utilized a quantile regression model to predict a probability distribution for the future citation count of a paper in the field of physics. Mazloumian (2012) investigated ten common but prominent citation indicators, and also employed a linear model to predict citation count. Abramo, D'Angelo, and Felici (2019) focused on analyzing a publication's early citations and the impact factor (IF) of the journal, and adopted two different linear models to predict citation count in the nine years after publication. They found that the IF of the journal becomes negligible two years after publication. Jimenez, Avila, Dueñas, and Gelbukh (2020) investigated the stylistic factors in the title and abstract of papers that may have an impact on citation count and presented a novel set of stylistic features. They achieved a mean absolute error of 0.805 with only the top-250 correlated features by using a linear model. Chakraborty, Kumar, Goyal, Ganguly, and Mukherjee (2014) examined over 1.5 million publications collected from the computer science field, and identified six broad categories of citation trajectories according to their rules. After that, they proposed a stratified learning framework where a paper is classified into one of six categories before predicting future citation count by support vector machine. In addition, Fu and Aliferis (2010), Robson and Mousquès (2016) and Wang, Yu, An, and Yu (2012), Wang, Yu, and Yu (2011) utilized support vector machine, decision tree, random forest, and K-nearest neighbor to predict citation count.

Due to its strong generalization ability, the neural network algorithm is also widely used in citation count prediction (Lu et al., 2021). Ruan et al. (2020) extracted some features of papers to predict five-year citations on a corpus of 49,834 papers in the library, information and documentation field, by using the feed-forward neural network model. In order to identify emerging technologies, Lee, Kwon, Kim, and Kwon (2018) employed 18 patent indicators and also employed a neural network model to predict the citation count of patents. Based only on the citation frequency in the first few years after publication, Abrishami and Aliakbary (2019) presented a "sequence to sequence" neural network model to predict long-term citations of a publication. Their experiments show that their proposed method outperforms state-of-the-art methods. Recently, some studies have shown that alternative metrics, often known as "altmetrics", provide new insights into the ways people consume, share, and report research, and have employed them in impact prediction (Barnes, 2015; Brody, Harnad, & Carr, 2006; Wooldridge & King, 2019). Akella et al. (2021) utilized altmetrics (e.g. social media shares, mentions) to predict long- and short-term citations. They tried various machine learning and deep learning models, and found that neural networks and ensemble models give the best performance for their research problem.

Compared with traditional citation count prediction, which only predicts citation frequency, FGCCP aims to predict in-text citation count from different structural functions, separately. To be specific, we treated it as a "sequence to sequence" issue, and the transformer (Vaswani et al., 2017), which has been proved to be an effective "sequence to sequence" model, was used as the basis of our proposed model to complete FGCCP.

## 3. Methodology

### 3.1. Problem definition

The goal of this study is to predict in-text citation count from each structural function of a paper, separately, which is referred to as a fine-grained citation count prediction task (FGCCP). Compared with traditional citation count prediction, FGCCP is more challenging in two sides. First, it not only predicts citation frequency, but also forecasts the structural function from which the additional citation count comes. Second, citations from different structural functions are not independent for a paper. For example, Introduction citations tend to also be cited within Background (Thelwall, 2019). Predicting citation count from one structural function may have a close connection with predicting citation count from another structural function. Hence, FGCCP, to some extent, belongs to a multi-task learning job.

In the following paragraphs, we clarify the FGCCP through Fig. 1. First, we identify the structural functions in a specific domain (i.e. $i_1, i_2, \ldots, i_5$), and count in-text citation frequency for papers. Subsequently, we treat FGCCP as a "sequence to sequence" issue. To be specific, let us suppose that a paper acquired $x_1^{(i)}, x_2^{(i)}, \ldots, x_S^{(i)}$ citations from a structural function, $i$, in $S$ consecutive years. The problem is to forecast citation count from the structural function, $i$, of the paper in the following $S'$ consecutive years. It is worth mentioning that the functional structure, $I$, may be composed of more than one structural function (Lu et al., 2018). For simplicity, we denoted citation count prediction in a structural function, $i$, as $FGCCP_i$ and all of $FGCCP_i$ constitutes FGCCP. Thus, FGCCP is fundamentally a fitting regression task, and each $FGCCP_i$ can be described in Eq. (1).

$$\left(\widehat{x_{S+1}^{(i)}}, \widehat{x_{S+2}^{(i)}}, \ldots, \widehat{x_{S+S'}^{(i)}}\right) = F^{(i)}\left(x_1^{(i)}, x_2^{(i)}, \ldots, x_S^{(i)}\right)\ i \in I \tag{1}$$

$x_t^{(i)}$ and $\widehat{x_t^{(i)}}$ denoted the actual value and predicted value, respectively. Notably, for convenience, we only utilized the in-text citations as the input and set $S$ equal to $S'$ in the following experiments.

In this paper, the transformer (Vaswani et al., 2017) is the basis of our proposed model. The transformer followed an encoder-decoder structure by using a stacked self-attention mechanism (Vaswani et al., 2017). Compared with common "sequence to sequence" models (e.g. LSTM and RNN (Elman, 1990; Hochreiter & Schmidhuber, 1997)), the transformer introduces positional encoding to record information about the relative position in time series data, and can be calculated in parallel. To accomplish FGCCP, we proposed the multi-transformers with among-attention mechanism (MTAT), which can leverage the correlation among $FGCCP_i$. The following section introduces our prediction model in detail.

### 3.2. Prediction model

In this study, the encoder-decoder structure, which has been illustrated as an effective way to solve a "sequence to sequence" problem, is employed. We proposed a transformer-based model, multi-transformers with among-attention mechanism (MTAT), to fulfill FGCCP. Specifically, for each structural function, $i$, we employed a transformer to fulfill $FGCCP_i$. As mentioned above, none of $FGCCP_i$ are independent of each other. Thus, we aggregated the loss function of each $FGCCP_i$ in the training phase, and proposed an among-attention mechanism in order to ensure the close connection between different $FGCCP_i$. Before introducing the among-attention mechanism, first, we briefly overview the model architecture of the transformer.

Let us start with Scaled Dot-Product Attention, which is the core module of the transformer. As shown in the left of Fig. 2, first, the

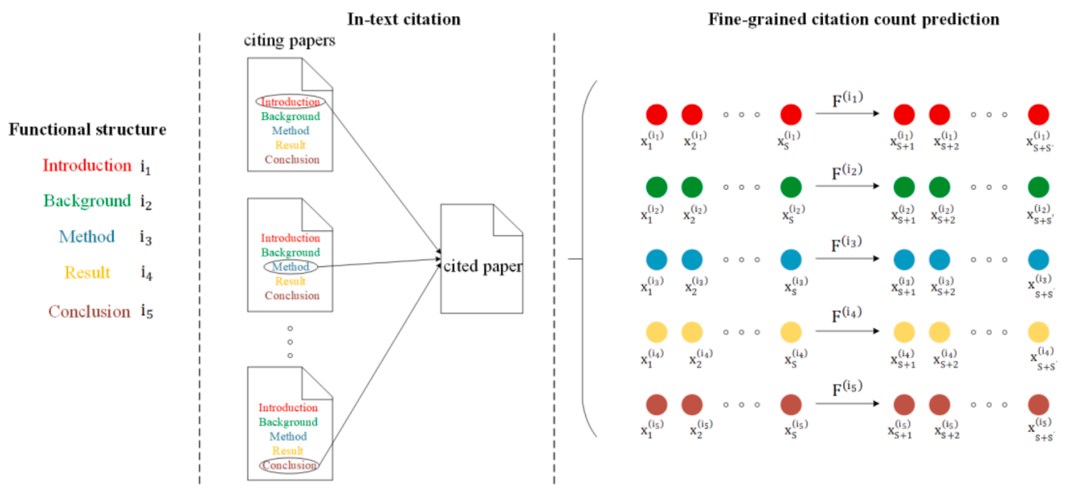



**Fig. 1.** The sketch map for clarifying fine-grained citation count prediction.

transformer maps the input sequence, $(x_1^{(i)}, x_2^{(i)}, \ldots, x_S^{(i)})$, into three matrices, $Q^{(i)}(i.e.(q_1^{(i)}, q_2^{(i)}, \ldots, q_S^{(i)})^T)$, $K^{(i)}(i.e.(k_1^{(i)}, k_2^{(i)}, \ldots, k_S^{(i)})^T)$, $V^{(i)}(i.e.(v_1^{(i)}, v_2^{(i)}, \ldots, v_S^{(i)})^T)$ by matrix operations which are as follows, Eqs. (2)-(4).

$$Q^{(i)^T} = W_Q^{(i)}(x_1^{(i)}, x_2^{(i)}, \ldots, x_S^{(i)}) \tag{2}$$

$$K^{(i)^T} = W_K^{(i)}(x_1^{(i)}, x_2^{(i)}, \ldots, x_S^{(i)}) \tag{3}$$

$$V^{(i)^T} = W_V^{(i)}(x_1^{(i)}, x_2^{(i)}, \ldots, x_S^{(i)}) \tag{4}$$

After that, the Scaled Dot-Product Attention module calculated the inner product between queries $Q^{(i)^T}$ and keys $K^{(i)}$ as weight, and computed a weighted sum of the values, $V^{(i)}$, as output $B^{(i)}$. The calculation formula and its corresponding matrix operation are shown in Eqs. (5) and (6), respectively.

$$b_s^{(i)^T} = softmax\left(q_s^{(i)^T} k_1^{(i)}, q_s^{(i)^T} k_2^{(i)}, \ldots, q_s^{(i)^T} k_S^{(i)}\right)\left(v_1^{(i)}, v_2^{(i)}, \ldots, v_S^{(i)}\right)^T s \in S \tag{5}$$

$$B^{(i)} = SelfAttention(Q^{(i)}, K^{(i)}, V^{(i)}) = softmax\left(\frac{Q^{(i)} K^{(i)^T}}{\sqrt{d_k}}\right) V^{(i)} \tag{6}$$

The overall model architecture structure of the transformer is shown in the middle of Fig. 2. The transformer employed Multi-Head Attention blocks (gray blocks), which are actually composed of multiple Scaled Dot-Product Attention in both encoder layer and decoder layer. The blue block is a simple feed-forward network (FFN), which is calculated by the following, Eq. (7). There also exists a residual connection and layer normalization in the transformer, and the circle represents positional encoding, which is the sine and cosine functions of different frequencies.

$$FFN(x) = W_2(ReLU(W_1 x + b_1)) + b_2 \tag{7}$$

In MTAT, for each $i \in I$, a transformer is employed to fit $F^{(i)}$. Hence, there are $|I|$ transformers in total. To make use of the correlation among $FGCCP_i$, we introduced the among-attention mechanism. The idea of the among-attention mechanism is very simple. In each

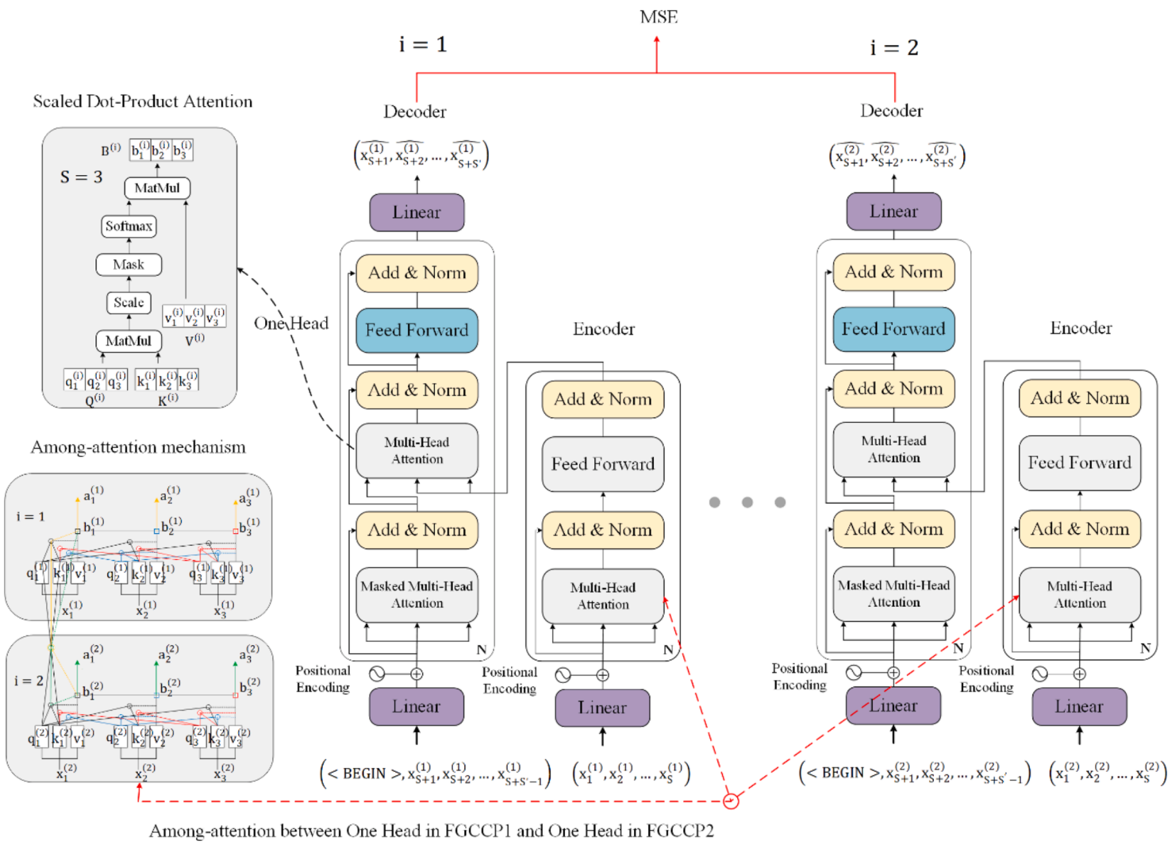



**Fig. 2.** Multi-transformers with among-attention mechanism (a case for $|I| = 2$ and $S = 3$).

time step of time series data, we also computed attention weight among queries from one transformer and keys from other transformers. After that, a weighted sum of outputs of Scaled Dot-Product Attention of $|I|$ transformers in that time step works as new outputs, $A^{(i)}$. In this study, we only utilized the among-attention mechanism among the Multi-Head Attention blocks of the encoder layer of the transformers. Actually, it can also be used among the Masked Multi-Head Attention blocks of the decoder layer of the transformers. In this way, citation history from a structural function, $i$, is not only used to fit $F^{(i)}$, but also to fit other $F^{(j)}$.

For the sake of understanding, we utilized $|I| = 2$ and $S = 3$ as a special case to clarify the among-attention mechanism, as shown in the left of Fig. 2. For the first input of time series data (i.e. $s = 1$), queries $(q_1^{(1)}, q_1^{(2)})$, keys $(k_1^{(1)}, k_1^{(2)})$, and outputs $(b_1^{(1)}, b_1^{(2)})$ of Scaled Dot-Product Attention of the encoder layers of two transformers can be computed. As shown in Eq. (8), we calculated inner product between $q_1^{(1)}$ and $(k_1^{(1)}, k_1^{(2)})$ followed by a *softmax* as weight, and computed the weighted sum of $b_1^{(1)}$ and $b_1^{(2)}$ as output, $a_1^{(1)}$. Similarly, we can obtain $a_1^{(2)}$. Both $a_1^{(1)}$ and $a_1^{(2)}$ possess citation information from two structural functions.

$$a_1^{(1)^T} = softmax\left(q_1^{(1)^T} k_1^{(1)}, q_1^{(1)^T} k_1^{(2)}\right)\left(b_1^{(1)}, b_1^{(2)}\right)^T \tag{8}$$

The general expression of the among-attention mechanism is as follows.

$$a_s^{(i)^T} = softmax\left(q_s^{(i)^T} k_s^{(1)}, q_s^{(i)^T} k_s^{(2)}, \ldots, q_s^{(i)^T} k_s^{(I)}\right)\left(b_s^{(1)}, b_s^{(2)}, \ldots, b_s^{(I)}\right)^T \; i \in I \text{ and } s \in S \tag{10}$$

The among-attention mechanism can also be calculated in the matrix operation, as shown in Eq. (11). $\otimes$ denotes element-wise product. The dimensions of $\widetilde{Q}(i.e.((Q^{(1)}, Q^{(2)}, \ldots, Q^{(I)})))$, $\widetilde{K}(i.e.(K^{(1)}, K^{(2)}, \ldots, K^{(I)}))$, and $\widetilde{V}(i.e.(V^{(1)}, V^{(2)}, \ldots, V^{(I)}))$ are all $(S, I, d_{model})$, where $d_{model}$ is determined by the dimension of $W_Q$. $h$ represents the number of heads in each transformer, and $d_k$ obeys $d_k = d_{model}/h$. In empirical experiments, before calculating the among-attention weight (i.e. $\widetilde{Q} \otimes \widetilde{K}^T$), the layer normalization is utilized to normalize the last dimension of $\widetilde{Q}$ and $\widetilde{K}$, respectively.

$$AmongAttention\left(\widetilde{Q}, \widetilde{K}, \widetilde{V}\right) = Softmax\left(\frac{\widetilde{Q} \otimes \widetilde{K}^T}{\sqrt{d_k}}\right) \otimes \widetilde{V} \tag{11}$$

The encoder-decoder structure of MTAT can be found in the right of Fig. 2. In the encoder layer of each transformer, $(x_1^{(i)}, x_2^{(i)}, \ldots, x_S^{(i)})$, which represents citations of a paper from a structural function, $i$, in $S$ consecutive years, is the input variable. In the decoder layer of each transformer, $(\langle BEGIN \rangle, x_{S+1}^{(i)}, x_{S+2}^{(i)}, \ldots, x_{S+S'-1}^{(i)})$ is the input variable during the training phase. $(\widehat{x_{S+1}^{(i)}}, \widehat{x_{S+2}^{(i)}}, \ldots, \widehat{x_{S+S'}^{(i)}})$ denotes the predicted value for citations of the paper from a structural function, $i$, in $S'$ consecutive years. In empirical experiments, $\langle BEGIN \rangle$ is set to 0.

### 3.3. Baselines

In this study, we selected seven baselines to evaluate and compare the prediction result. First, the “Mean of Early Years” method (MEY) proposed by Abrishami and Aliakbary (2019) is employed as a baseline. MEY is a simple but effective prediction method which always uses the average of citation count in sequence $S$ to predict citation count in sequence $S'$, and presents a relatively satisfactory prediction performance in some case. We employed MEY to separately predict citation count from each structural function, $i$, and the mathematical formula of MEY is as follows.

$$\widehat{x_{S+1}^{(i)}} = \widehat{x_{S+2}^{(i)}} = \ldots = \widehat{x_{S+S'}^{(i)}} = \frac{1}{S}\sum_{s=1}^{S} x_s^{(i)} \; i \in I \tag{18}$$

Recent studies showed that machine learning (ML) algorithms have achieved excellent prediction performance on citation count prediction (Abramo et al., 2019; Bütün, Kaya, & Alhajj, 2017; Chakraborty et al., 2014; Djokoto et al., 2020; Fu & Aliferis, 2010; Yan, Huang, Tang, Zhang, & Li, 2012). This study employed three common machine learning approaches (i.e. linear model (LR), random forest (RF), eXtreme Gradient Boosting (XGBoost)) as baselines. Specifically, in order to fulfill the “sequence to sequence” problem considered in this paper, an auto-regression method is used in each baseline approach during the training and testing phases. The general mathematical expression is as follows.

$$\widehat{x_{t+S+1}^{(i)}} = F_{ML}^{(i)}\left(x_{t+1}^{(i)}, x_{t+2}^{(i)}, \ldots, x_{t+S}^{(i)}\right) \tag{18a}$$

$0 \le t \le S' - 1$ and $i \in I$where $F_{ML}^{(i)}$ represents a machine learning model for the $FGCCP_i$. The citation frequency of a paper in a continuous time series of length $S$ is adopted to forecast the citation frequency of the paper at the next moment. Notably, in the training phase, we always used the true value as the input, which is based on the teacher forcing mechanism. In the testing phase, the predicted value of the previous stage is utilized as the input for the next prediction. For example, in the testing phase, we first utilized $(x_1^{(i)}, x_2^{(i)}, x_3^{(i)})$ to predict $\widehat{x_4^{(i)}}$, then $(x_2^{(i)}, x_3^{(i)}, \widehat{x_4^{(i)}})$ to predict $\widehat{x_5^{(i)}}$, and so on.

The deep learning technology shows extraordinary talent in a variety of prediction tasks, and the citation count prediction is no exception (Abrishami & Aliakbary, 2019; Akella et al., 2021; Lee et al., 2018; Ruan et al., 2020). In this study, we also employed LSTM and RNN, which are the two most basic but very successful time series prediction models, as baselines. The model architecture of LSTM and RNN deployed in this paper is shown in Fig. 3. Moreover, to analyze the utility of our proposed among-attention mechanism, we excluded the among-attention mechanism in our proposed model as the last baseline.

### 3.4. Evaluation

We employed three common metrics to evaluate the prediction performance of our proposed method and baselines. Specifically, mean average error (*MAE*), mean square error (*MSE*), and the coefficient of determination ($R^2$) are utilized in this paper. *MAE* and *MSE* gage the average of absolute errors between predicted and actual values and the variation of the predicted values to the actual values, respectively. $R^2$ measures the proportion of total variance explained by the model. Therefore, the smaller *MAE* as well as *MSE*, and the larger $R^2$ are preferred. Eqs. (19)-(21) illustrate *MAE*, *MSE*, and $R^2$, respectively.

$$MAE_{total} = \frac{1}{N}\sum_{n}^{N}\sum_{i}^{I}\sum_{s=S+1}^{S+S'}\left|\widehat{x_{n,s}^{(i)}} - x_{n,s}^{(i)}\right| \tag{19}$$

$$MSE_{total} = \frac{1}{N}\sum_{n}^{N}\sum_{i}^{I}\sum_{s=S+1}^{S+S'}\left(\widehat{x_{n,s}^{(i)}} - x_{n,s}^{(i)}\right)^2 \tag{20}$$

$$R_{total}^2 = \frac{\sum_{n}^{N}\sum_{i}^{I}\sum_{s=S+1}^{S+S'}\left(\widehat{x_{n,s}^{(i)}} - x_{n,s}^{(i)}\right)^2}{\sum_{n}^{N}\sum_{i}^{I}\sum_{s=S+1}^{S+S'}\left(\overline{x_{n,s}^{(i)}} - x_{n,s}^{(i)}\right)^2} \tag{21}$$

In the above formulas, $\widehat{x_{n,s}^{(i)}}$ and $x_{n,s}^{(i)}$ represent the predicted value and actual value, respectively. *N* indicates the sample sizes. *I*, *S*, and $S'$ have the same meanings as stated above. During the training phase, $MSE_{total}$, which aggregates loss from all $FGCCP_i$, is utilized as the final loss function. In addition, *MAE*, *MSE*, and $R^2$ on the $FGCCP_i$ are denoted as $MAE_i$, $MSE_i$, and $R_i^2$, respectively.

## 4. Data

### 4.1. Data collection and preprocessing

There are many publicly available scholarly data sets, which provides structured full text, such as S2ORC (Lo, Wang, Neumann, Kinney, & Weld, 2019) and unarXive (Saier & Färber, 2020). In this study, we chose research articles from PubMed Central Open Access Subset (PMC—OAS) as our data set (hereafter, PMC dataset), which are well-structured documents in the biomedical field. That being said, there are relatively standardized section headers of articles, and structural functions can be easily identified. As of May 2020, we collected full-text documents from PubMed Central (PMC) Open Access Subset comprising 1479,688 research articles. The publication distribution in the current dataset is shown in Fig. 4. We parsed these XML documents to obtain the metadata of the articles and the citation locations of their references. After that, we de-duplicated the collected research articles and extracted references according to article title, author(s), and publication year. Finally, we obtained 68,130,759 pieces of citation context data, and 11,424, 332 unique papers.

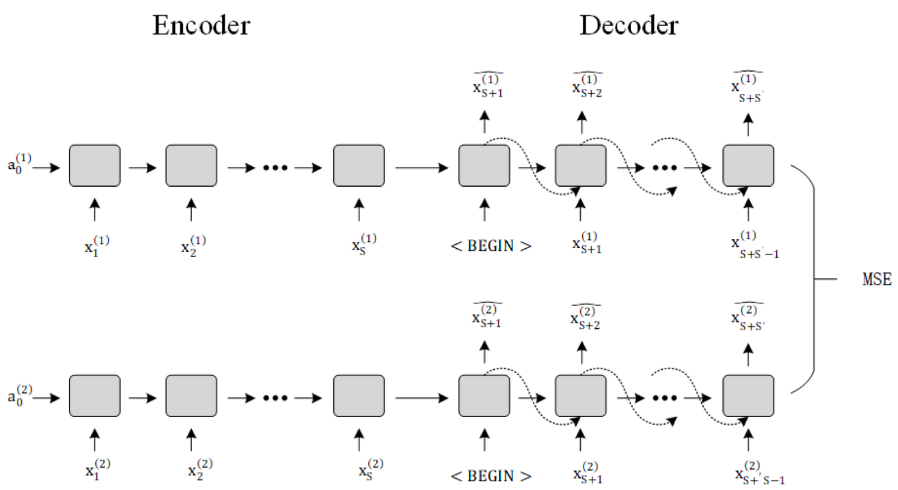



**Fig. 3.** LSTM and RNN framework.

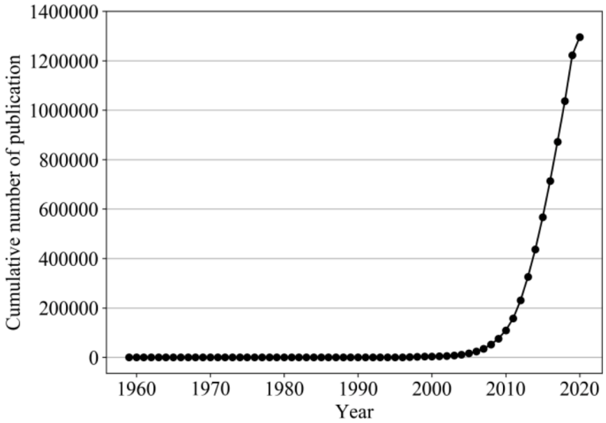



**Fig. 4.** Distribution of publications.

In this study, we used the functional structure clustering algorithm presented by Lu et al. (2018) to identify functional structure and structural functions in the collected biomedical literatures. To be specific, we extracted the section header, which is at the nearest front of citation location, from the citation context data. After that, we converted these section headers to lowercase letters, and removed their punctuation and number. Subsequently, these section headers were ranked in descending order of frequency. *Top*10 section headers account for 91.45% of total citation context data, while *Top*100 section headers account for 92.14%. For simplicity, only *Top*10 high-frequency section headers are selected to identify five structural functions (i.e. Introduction, Background, Method, Experiment and result (hereafter, Experiment), and Discussion and conclusion (hereafter, Discussion)). The clustering results are listed in Table 1 in detail. We excluded citation context data (8.55%) whose section headers are not in the *Top*10, and used the remaining 91.45% pieces of citation context data to count citations for unique papers. Notably, the full counting method (Pak et al., 2020) was utilized to count citation frequency in each structural function separately. Distribution of structural functions on the remaining 62,302,506 items of citation context data can be found in Table 2.

### 4.2. Training set

It is empirically observable that citation distributions generally are heavy-tailed distributed (Bu et al., 2021; Huang, Bu, Ding, & Lu, 2020; Onodera & Yoshikane, 2015; Perianes-Rodriguez & Ruiz-Castillo, 2016); that is, a small number of publications are highly cited, while most of them are lowly cited. To construct a balanced dataset and avoid too many meaningless zero values in our input and output data, we chose $x_{min}$ and $x_{max}$ to classified articles into three categories. As shown in Table 3, $x_{min}$ and $x_{max}$ are manually selected by observing the citation distribution, and are set to 10 and 100, respectively. Subsequently, we eliminated papers with fewer than ten citations, and randomly selected 30,000 papers from the citation frequency range of $[10, 100)$, and retained 29,390 papers with a citation count of more than 100. Finally, a subset of 59,390 papers was obtained. Previous studies showed that citations of a paper within five years after publication is an important reflection of its quality (Wang et al., 2011, 2012), and employed 3-to-5-year citation time window to predict its citations in a long-term period (Abramo et al., 2019; Abrishami & Aliakbary, 2019; Ruan et al., 2020). In this paper, for simplicity, $S$ and $S'$ are both set to 5.We employed the sliding window method (Lu et al., 2021; Xu, Hao, An, Yang, & Wang, 2019) to generate the input-output pairs. Specifically, for each paper from the subset, we slid the citation time window in turn to generate the input-output pairs of length, $S + S'$, starting from the year of publication of the paper. Hence, each input-output pair for $FGCCP_i$ is in the form of a tuple, $((x_1^{(i)}, x_2^{(i)}, \dots, x_S^{(i)}), (x_{S+1}^{(i)}, x_{S+2}^{(i)}, \dots, x_{S+S'}^{(i)}))$. Finally, we split all the input-output pairs into train data and test data at a ratio of 8:2, and the sample sizes of the train and test data are shown in Table 4.

**Table 1**
The functional structure schema generated from the PMC dataset.

| Functional structure | Section headers |
|---|---|
| Introduction | introduction |
| Background | background, related work |
| Method | method |
| Experiment and result | result, experimental section, experimental procedure, case |
| Discussion and conclusion | discussion, conclusion |

**Table 2**
Distribution of structural functions on the PMC dataset.

| Functional structure | Frequency |
|---|---|
| Introduction | 19,799,902 |
| Background | 2,667,109 |
| Method | 8,410,505 |
| Experiment and result | 9,529,914 |
| Discussion and conclusion | 21,895,076 |

**Table 3**
Citation count distribution.

| Interval | $[0,\ 10)$ | $[10, 100)$ | $[100, +\infty)$ | Total |
|---|---|---|---|---|
| Number of publications | 9,978,799 | 1,416,143 | 29,390 | 11,424,332 |

**Table 4**
Sample size in train and test data.

| Input-output pairs | Sample size |
|---|---|
| Train data | 283,290 |
| Test data | 70,823 |
| Total | 354,113 |

## 5. Experiments

### *5.1. Parameters*

According to the $MSE_{total}$ on the test set, we chose the optimal setting of the model architecture and training parameter. As shown in Table 5, the number of layers in both Encoder and Decoder is set to 1. $d_{model}$ and the number of heads in each transformer, $h$, are set to 16 and 2, respectively. Therefore, $d_k$, which obeys $d_k = d_{model}/h$, is 8. In addition, the dimension of the output of FFN, $d_{ffn}$, is 64. To prevent an overfitting problem, dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) with a ratio of 0.1 is applied to the output of each sub-layer. Finally, the total number of parameters of our proposed model is about $3.9 \times 10^4$. The training parameters can be found in the last four rows of Table 5. The parameters of the Adam optimizer are the same as the settings adopted by Vaswani et al. (2017).

LR, RF, and XGB were implemented through the algorithm library encapsulated in scikit-learn (Pedregosa et al., 2011), and the grid search algorithm was used to determine the parameters of RF and XGB. The model architectures of LSTM and RNN are similar to that of MTAT. The number of layers in the encoder and decoder of LSTM and RNN is set to 1, and hidden units are set to 32 and 64, respectively. The total parameters of both LSTM and RNN are about $4.3 \times 10^4$. We also tried a variety of compositions of hyper parameters, but they did not significantly improve the prediction performance of LSTM and RNN. In the end, all neural network algorithms are implemented by the "TensorFlow" framework.

**Table 5**
Parameters of multi transformers with among-attention mechanism.

| Parameter | Value |
|---|---|
| Number of layers in Encoder | 1 |
| Number of layers in Decoder | 1 |
| $d_{model}$ | 16 |
| $h$ | 2 |
| $d_k$ | 8 |
| $d_v$ | 8 |
| $d_{ffn}$ | 64 |
| $P_{drop}$ | 0.1 |
| Epochs | 25 |
| Batch size | 64 |
| Initial learning rate | 1e-6 |
| Optimizer | Adam |

### 5.2. Results

During the process of training the neural network, the random gradient descent method was utilized to update the parameters of the model along the negative gradient direction. The $MSE_{total}$ variations of our proposed model on the training set can be found in Fig. 5. The model parameters tend to converge after ten epochs, and the $MSE_{total}$ decreases slightly in the remaining fifteen epochs. Finally, the $MSE_{total}$ of optimal MTAT on the training set, which is calculated based on teacher forcing mechanism during training process, is 9.4675.

To evaluate the prediction performance of MTAT, in the following paper, we employed the auto-regression method mentioned in "Baselines" subsection. $MSE_{total}$, $MAE_{total}$, and $R^2_{total}$ on the training set are computed, and are 16.1359, 1.1021, and 0.7696, respectively. $MSE_{total}$, $MAE_{total}$, and $R^2_{total}$ on the test set are calculated, and are 15.0693, 1.1311, and 0.7133, respectively. As shown in Table 6, we also calculated the above three criteria on the test set for different $FGCCP_i$. We found that our proposed model performs best on predicting citations from Method, and performs worst on predicting citations from Background in terms of $R^2$. The $R^2$ on $FGCCP_{Method}$ and $FGCCP_{Background}$ (i.e. $R^2_{Method}$ and $R^2_{Background}$) are 0.7537 and 0.2671, respectively. In addition, the $R^2_{Introduction}$, $R^2_{Experiment}$, and $R^2_{Discussion}$ show that our model achieved a satisfied prediction accuracy on the remaining $FGCCP_i$. However, in terms of $MSE$ and $MAE$, our model obtains the best prediction performance on $FGCCP_{Background}$ and the worst prediction performance on $FGCCP_{Method}$. This is caused by the fact that our collected citation context data possesses the minimal number of citations from Background, as shown in Table 2. Hence, it is difficult for the model to perform well on $FGCCP_{Background}$, and obtain a lower $R^2$. However, on the other side, the minimal number of citations from Background results in a lot of zero values in input-output pairs of $FGCCP_{Background}$. Thus, it is easy to make $MSE_{Background}$ and $MAE_{Background}$ smaller. We also utilized citation distribution on all the input-output pairs to clearly clarify our point. As shown in Fig. 6, the x-axis represents the natural logarithm of citation frequency (i.e. $log(\sum_{s}^{S+S'} x_s^{(i)} + 1)$), and the y-axis indicates normalized frequency, which means distribution density (Fig. 6 is plotted by seaborn module (Waskom, 2021)). Although we have excluded papers with fewer than ten citations during balanced dataset construction, there are still many zero values in all the input-output pairs, and this phenomenon is most obvious in input-output pairs from Background. In addition, we also set $x_{min}$ as 20 and other values for $S$ and $S'$ (e.g., 3 and 4), and repeated the above experiments. The prediction performance of our proposed model is almost unaffected.

In order to further analyze the prediction performance of our proposed model, we also randomly selected twelve samples from the test set as a case study. As shown in Fig. 7(a)-(l), the x-axis and y-axis denote time and citation frequency, respectively. The gray area represents the input sequence, and the length of the input sequence is $S$. The solid line indicates the actual value and the dotted line represents the predicted value, and the predicted and actual trajectories of different $FGCCP_i$ are shown through curves in different colors. We found that, despite some deviations, the predicted value properly reflects the variation of the real value in the following $S'$ consecutive years. Our proposed model achieves satisfactory prediction performance on FGCCP.

To compare the prediction accuracy of our proposed model and baselines, we first randomly divided all the input-output pairs generated by the balanced dataset ten times at the ratio of 8: 2. After that, ten groups of training set and test set were obtained. MEY can directly complete the prediction on the test set. Machine learning and deep learning methods were trained on the training set, after which $MAE$, $MSE$, and $R^2$ on the test set were calculated. The average of each criterion on the ten groups of test set for the different models is listed in Table 7. In addition, the paired t-tests are employed to test the significant difference between MTAT and baselines. MTAT achieves a better prediction performance in terms of all criteria than MEY, LR, RF, XGB, and multi-transformers without among-attention mechanism (hereafter, MT). Specifically, the $MSE$ of MTAT is lower than 41.60% for MEY, 14.45% for LR, 42.87% for RF,

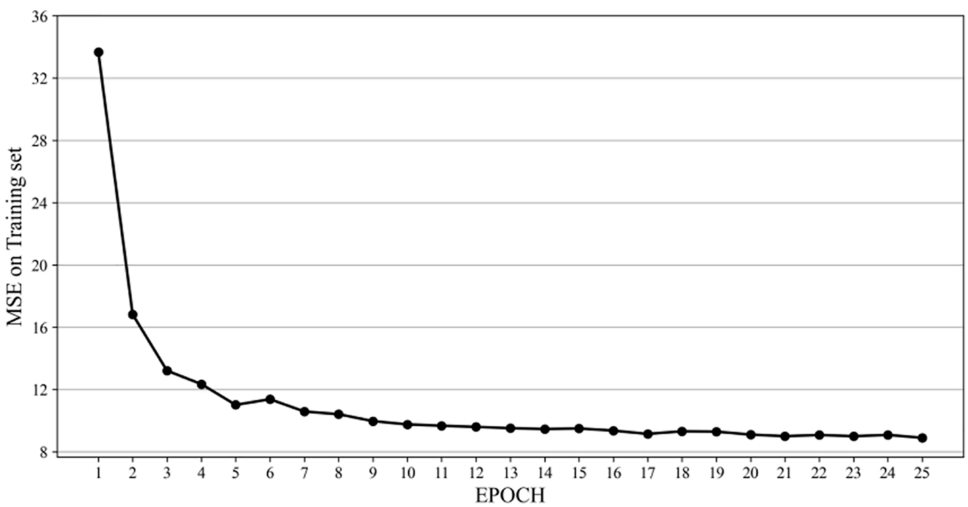



**Fig. 5.** $MSE_{total}$ during the training process.

**Table 6**
*MAE*, *MSE*, and $R^2$ on the test set.

| Structural function | *MAE* | *MSE* | $R^2$ |
|---|---|---|---|
| Introduction | 1.5457 | 11.7414 | 0.6249 |
| Background | **0.2684** | **0.8047** | 0.2671 |
| Method | 1.5994 | 49.5885 | **0.7537** |
| Experiment and result | 0.9285 | 6.8025 | 0.4595 |
| Discussion and conclusion | 1.3135 | 6.4096 | 0.5218 |
| Total | 1.1311 | 15.0693 | 0.7133 |

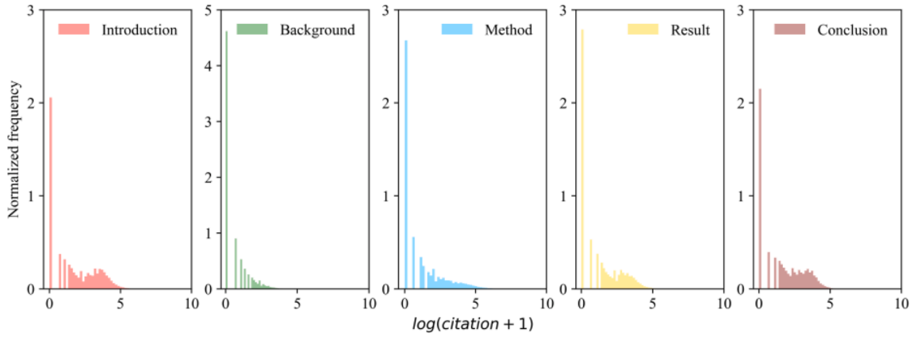



**Fig. 6.** Citation distribution on the input-output pairs.

12.84% for XGB, and 15.04% for MT. The *MAE* of MTAT is lower than 7.96% for MEY, 30.61% for LR, 19.36% for RF, 16.80% for XGB, and 4.39% for MT. In addition, p-value shows that there exists a significant difference between MTAT and MEY, LR, RF, XGB, and MT in all criteria. As for RNN and LSTM, MTAT achieves a better prediction accuracy in term of *MAE*. The *MAE* of MTAT is lower than 12.69% for RNN, and 13.04% for LSTM, and the difference is statistically significant. However, RNN, LSTM, and MTAT have no significant difference in term of *MSE* and $R^2$. Notably, compared with RNN and LSTM, the transformer can be calculated in parallel (Vaswani et al., 2017), and trained faster on the premise of a similar magnitude of number of parameters. In addition, the total number of parameters of RNN, LSTM, and MTAT deployed in this study is similar. Thus, the current comparison is fair.

In order to analyze the effect of the among-attention mechanism on the prediction performance of FGCCP, in Table 8, we report the prediction result of MTAT and MT in detail. As discussed previously, three metrics are still the average on the ten groups of the test set. We found that, for all $FGCCP_i$, MTAT obtained better prediction accuracy than MT in terms of all criteria, which demonstrates the importance of the among-attention mechanism. In the following subsection, we further analyze the role of the among-attention mechanism.

### 5.3. Analyzing the role of the among-attention mechanism

To figure out potential reasons why the among-attention mechanism works, we also employed the heat map to analyze its role in MTAT. Specifically, the among-attention weights of Fig. 7(k) (l) are shown in Fig. 8(a) (b) as a case study. As we mentioned above, the among-attention mechanism works in each time step among the encoder layer of different $FGCCP_i$, and two heads are employed in our model. Hence, there are ten weight matrices for each case. Each of these matrices depicts the attention weight among one head of different $FGCCP_i$ at one time step. Therefore, the dimensions of each matrix are $I \times I$ (i.e. $5 \times 5$). The block in dark color indicates a lager weight and the block in light color is the opposite. Similar to Fig. 1, $i_1, i_2, \ldots$and, $i_5$ indicates Introduction, Background, Method, Experiments, and Results, respectively. In Head 1 of Fig. 8(a) (b), we found that it is easier to obtain higher weights for $(i_1, i_2)$ and $(i_3, i_4)$. Indeed, Introduction and Background may both play the role of introducing the research background, and Introduction citations tend to also be cited within Background (Thelwall, 2019). Moreover, Method and Result are bound to be closely linked. Hence, the model may learn to grasp the logical connections among citations from different structural functions, and use this potential relationship to improve the prediction performance of FGCCP. In Head 2 of Fig. 8(a) (b), the among-attention mechanism assigns a bigger weight to the $FGCCP_i$ which possesses more citations in $S$. For example, in Time 1 and 2, Head 2 of both cases assigns more weight to $i_1$, where citation information is abundant, than to citations from other structural functions. This might mean that our model flexibly used sufficient citation information from one structural function to predict citation frequency from other structural function where citation information is scarce. Therefore, our model may successfully find the internal connections between different but highly relevant tasks (e.g. $FGCCP_i$), so as to improve the overall prediction performance.

## 6. Discussion

In this study, we proposed a novel fine-grained citation count prediction task (FGCCP) which aims to predict in-text citation count

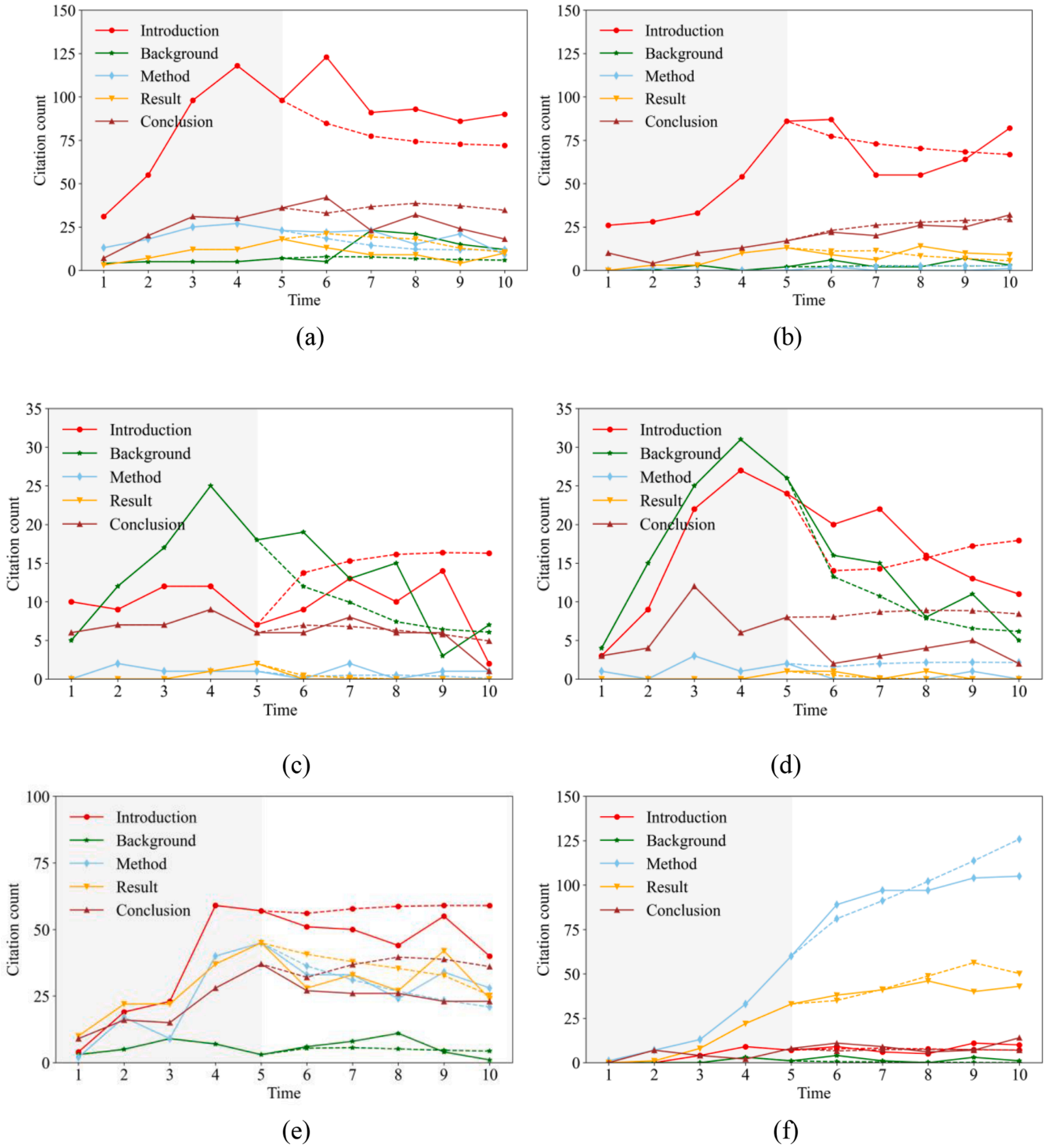



**Fig. 7.** Twelve cases from the test set.

from different structural functions separately. We also presented a transformer-based model (i.e. MTAT) in which a novel among-attention mechanism is employed. Based on a case study of research articles collected from PubMed Central Open Access Subset, we found that our model achieves satisfactory prediction accuracy and exceeds common machine learning and deep learning models on FGCCP.

### 6.1. *Theoretical implications*

This study has the following theoretical implications. First, we proposed FGCCP, which predicts not only citation frequency but also citation location. Compared with traditional citation count prediction, FGCCP aims to predict accurately the dynamics of in-text citation frequency from different structural functions of a paper. Therefore, for academic institutions, technology companies, and government bodies, FGCCP may provide more detailed decision-making evidence and evaluation basis in scientific research evaluation. Second, we proposed MTAT, which is a transformer-based model and takes all the advantages of the transformer. For example, MTAT can also be calculated in parallel, unlike other "sequence to sequence" models like RNN and LSTM. Most importantly, in MTAT,

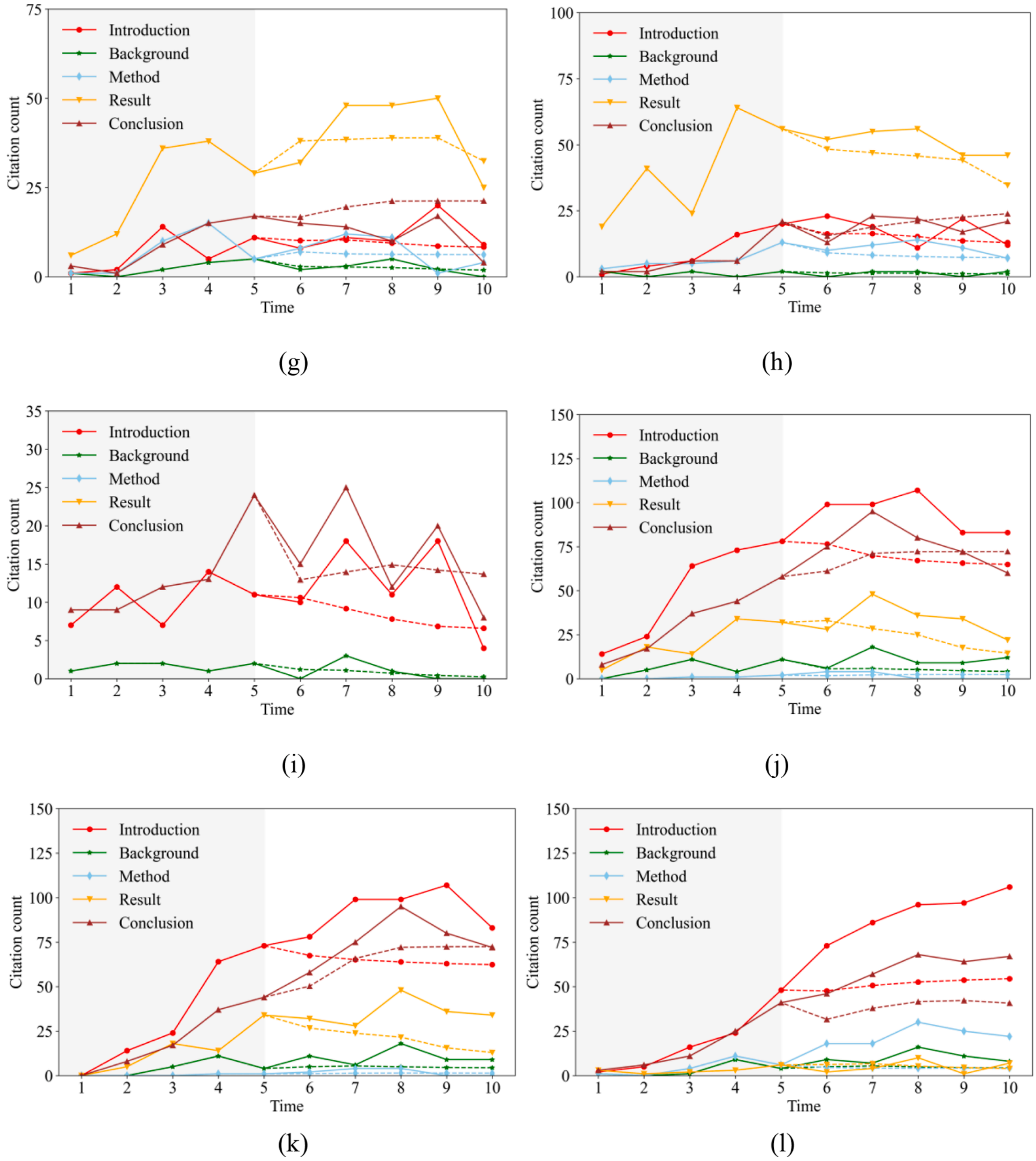



**Fig. 7.** (*continued*).

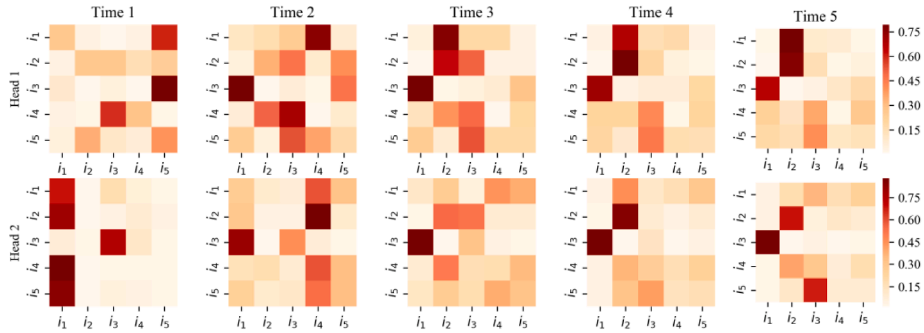
**Table 7**
Prediction performance on test set for different models.

| Model | *MAE* | *MSE* | $R^2$ |
|---|---|---|---|
| MEY | 1.1796*** | 26.8120*** | 0.4868*** |
| LR | 1.5646*** | 18.3054*** | 0.6486*** |
| RF | 1.3463*** | 27.4106*** | 0.4733*** |
| XGB | 1.3049*** | 17.9681*** | 0.6556*** |
| RNN | 1.2432*** | 15.7761 | 0.6979 |
| LSTM | 1.2485*** | **15.5455** | **0.7031** |
| MT | 1.1355*** | 18.4329*** | 0.6474*** |
| MTAT | **1.0857** | 15.6609 | 0.7006 |

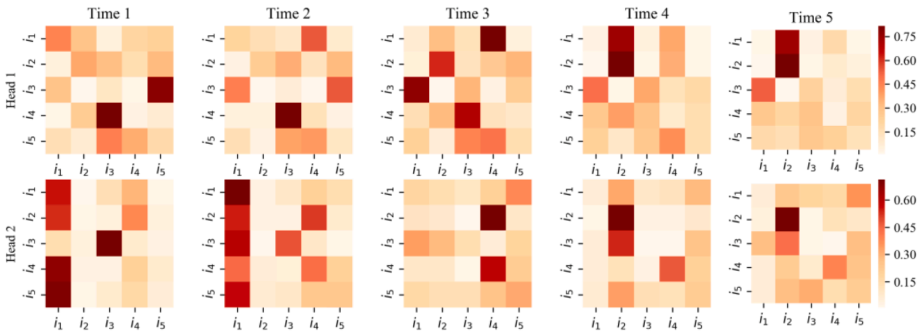
Note: * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$.

**Table 8**
Prediction performance on test set.

| Structural function | MT | | | MTAT | | |
|---|---|---|---|---|---|---|
| | *MAE* | *MSE* | $R^2$ | *MAE* | *MSE* | $R^2$ |
| Introduction | 1.5126 | 13.1655 | 0.5280 | 1.4878 | 11.7800 | 0.5767 |
| Background | 0.2739 | 0.8127 | 0.2490 | 0.2679 | 0.7915 | 0.2683 |
| Method | 1.6345 | 64.0402 | 0.6858 | 1.4852 | 52.2713 | 0.7438 |
| Experiment and result | 0.9823 | 7.2253 | 0.4200 | 0.9365 | 6.6804 | 0.4539 |
| Discussion and conclusion | 1.2745 | 6.9208 | 0.4829 | 1.2508 | 6.6580 | 0.5026 |
| Total | 1.1355 | 18.4329 | 0.6474 | 1.0857 | 15.6609 | 0.7006 |




(a) The among-attention weight of Fig. 7 (k)




(b) The among-attention weight of Fig. 7 (l)

**Fig. 8.** Two cases for analyzing the among-attention weight.

a simple but effective among-attention mechanism is utilized. The mechanism does not attach additional parameters except those required by layer normalization. However, it can obviously improve the prediction performance. After analyzing the among-attention weight, we found that the among-attention mechanism may help the model to understand the internal logical relationship among different $FGCCP_i$. Finally, based on a case study of the PMC dataset, MTAT achieves satisfactory prediction accuracy, and surpasses other common machine learning and deep learning models on FGCCP. This further proves the superiority of our model.

### *6.2. Practical implications*

Previous studies have suggested that citations should not be weighted equally, and citation mention and citation location suggest different contributions of the cited articles (Ding et al., 2013; Hu et al., 2013; Lu et al., 2017). Due to the fact that citations from different sections may be different in their citing motivation, and may reflect the types of cited papers (Hu et al., 2013; Tahamtan & Bornmann, 2018), this paper employed the definition of structural functions proposed by Lu et al. (2018) to determine the citation location reasonably. Thus, FGCCP may be combined with other content-based citation analysis research to carry out scientific research evaluation in a more reasonable way. For example, by weighting citations or filtering citations from different structural functions, just as Zhao and Strotmann (2020) did, FGCCP can differentiate the scientific impact of publications even when they have roughly the same citation frequency. Moreover, the MTAT presented in this study is a general model which can be easily deployed in other multi-task learning jobs.

## 7. Limitations

There are still some limitations in this study. Firstly, in this study, we only employed the in-text citation count from different structural functions as the input variable of FGCCP to keep the problem simple and general. Many studies have confirmed that other relevant features of scientific papers and altmetrics can significantly improve the performance of citation count prediction (Akella et al., 2021; Onodera & Yoshikane, 2015; Ruan et al., 2020; Yu et al., 2014). Hence, whether these aforementioned features can significantly improve the prediction accuracy of FGCCP needs to be further explored. Secondly, we only fulfilled FGCCP on the collected PMC dataset, in which the structural layout of publications (e.g. section headers) is fairly standardized. However, in other areas, whether our proposed model can still achieve satisfactory prediction performance on FGCCP also requires further study, especially in some cross-discipline bibliographic databases.

## CRediT authorship contribution statement

**Shengzhi Huang:** Conceptualization, Methodology, Formal analysis. **Yong Huang:** Conceptualization, Methodology, Writing – original draft. **Yi Bu:** Writing – original draft, Data curation. **Wei Lu:** Supervision, Formal analysis. **Jiajia Qian:** Investigation, Data curation. **Dan Wang:** Investigation, Writing – original draft, Writing – review & editing.

## Declaration of Competing Interest

The authors declare no competing interests.

## Acknowledgments

This work was supported by the Youth Science Foundation of the National Natural Science Foundation of China (grant no. 72004168).

## References

Abramo, G., D'Angelo, C. A., & Felici, G. (2019). Predicting publication long-term impact through a combination of early citations and journal impact factor. *Journal of Informetrics, 13*, 32–49.
Abrishami, A., & Aliakbary, S. (2019). Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics, 13*, 485–499.
Akella, A. P., Alhoori, H., Kondamudi, P. R., Freeman, C., & Zhou, H. (2021). Early indicators of scientific impact: Predicting citations with altmetrics. *Journal of Informetrics, 15*, Article 101128.
Bai, X., Zhang, F., & Lee, I. (2019). Predicting the citations of scholarly paper. *Journal of Informetrics, 13*, 407–418.
Barnes, C. (2015). The use of altmetrics as a tool for measuring research impact. *Australian Academic & Research Libraries, 46*, 121–134.
Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics, 12*, 59–73.
Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology, 57*, 1060–1072.
Bu, Y., Lu, W., Wu, Y., Chen, H., & Huang, Y. (2021). How wide is the citation impact of scientific publications? A cross-discipline and large-scale analysis. *Information Processing & Management, 58*, Article 102429.
Burrell, Q. L. (2002). Will this paper ever be cited? *Journal of the American Society for Information Science and Technology, 53*, 232–235.
Burrell, Q. L. (2003). Predicting future citation behavior. *Journal of the American Society for Information Science and Technology, 54*, 372–378.
Bütün, E., Kaya, M., & Alhajj, R. (2017). A supervised learning method for prediction citation count of scientists in citation networks. In *2017 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 952–958). IEEE.
Cao, X., Chen, Y., & Liu, K. R. (2016). A data analytic approach to quantifying scientific impact. *Journal of Informetrics, 10*, 471–484.
Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2014). Towards a stratified learning approach to predict future citation counts. In *IEEE/ACM joint conference on digital libraries* (pp. 351–360).
Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics, 7*, 583–592.
Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology, 65*, 1820–1833.
Djokoto, J. G., Agyei-Henaku, K. A. A., Afrane-Arthur, A. A., Badu-Prah, C., Gidiglo, F. K., & Srofenyoh, F. Y. (2020). What drives citations of frontier application publications? *Heliyon, 6*, e05428.
Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., & Radev, D. (2008). Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology, 59*, 51–62.
Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.
Fu, L., & Aliferis, C. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics, 85*, 257–270.
Glänzel, W., & Schubert, A. (1995). Predictive aspects of a stochastic model for citation processes. *Information Processing & Management, 31*, 69–80.
Herlach, G. (1978). Can retrieval of information from citation indexes be simplified? Multiple mention of a reference as a characteristic of the link between cited and citing article. *Journal of the American Society for Information Science, 29*, 308–310.
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735–1780.
Hu, Z., Chen, C., & Liu, Z. (2013). Where are citations located in the body of scientific articles? A study of the distributions of citation locations. *Journal of Informetrics, 7*, 887–896.
Huang, Y., Bu, Y., Ding, Y., & Lu, W. (2020). Partitioning highly, medium and lowly cited publications. *Journal of Information Science*, Article 0165551520917655.
Jimenez, S., Avila, Y., Dueñas, G., & Gelbukh, A. (2020). Automatic prediction of citability of scientific articles by stylometry of their titles and abstracts. *Scientometrics, 125*, 3187–3232.

Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change, 127*, 291–303.
Lo, K., Wang, L. L., .Neumann, M., Kinney, R., & Weld, D. S. (.2019). S2ORC: The semantic scholar open research corpus. arXiv preprint arXiv:1911.02782.
Lu, C., Ding, Y., & Zhang, C. (2017). Understanding the impact change of a highly cited article: A content-based citation analysis. *Scientometrics, 112*, 927–945.
Lu, W., Huang, S., Yang, J., Bu, Y., Cheng, Q., & Huang, Y. (2021). Detecting research topic trends by author-defined keyword frequency. *Information Processing & Management, 58*, Article 102594.
Lu, Wei, Huang, Yong, Bu, Yi, & Cheng, Qikai (2018). Functional structure identification of scientific documents in computer science. *Scientometrics, 115*, 463–486.
Mazloumian, A. (2012). Predicting scholars' scientific impact. *PloS One, 7*, e49246.
Mingers, J., & Burrell, Q. L. (2006). Modeling citation behavior in management science journals. *Information Processing & Management, 42*, 1451–1464.
Onodera, N., & Yoshikane, F. (2015). Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology, 66*, 739–764.
Pak, C. M., Wang, W., & Yu, G. (2020). An analysis of in-text citations based on fractional counting. *Journal of Informetrics, 14*, Article 101070.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825–2830.
Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2016). University citation distributions. *Journal of the Association for Information Science and Technology, 67*, 2790–2804.
Robson, B. J., & Mousquès, A. (2016). Can we predict citation counts of environmental modelling papers? Fourteen bibliographic and categorical variables predict less than 30% of the variability in citation counts. *Environmental Modelling & Software, 75*, 94–104.
Ruan, X., Zhu, Y., Li, J., & Cheng, Y. (2020). Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics, 14*, Article 101039.
Saier, T., & Färber, M. (2020). unarXive: A large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics, 125*, 3085–3108.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The journal of Machine Learning Research, 15*, 1929–1958.
Stegehuis, C., Litvak, N., & Waltman, L. (2015). Predicting the long-term citation impact of recent publications. *Journal of Informetrics, 9*, 642–657.
Suppe, F. (1998). The structure of a scientific paper. *Philosophy of Science, 65*, 381–405.
Tahamtan, I., & Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics, 12*, 203–216.
Thelwall, M. (2019). Should citations be counted separately from each originating section? *Journal of Informetrics, 13*, 658–678.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. et al. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.
Voos, H., & Dagaev, K. S. (1976). Are all citations equal? Or, Did We Op. Cit. Your Idem? *Journal of Academic Librarianship, 1*, 19–21.
Wan, X., & Liu, F. (2014). Are all literature citations equally important? Automatic citation strength estimation and its applications. *Journal of the Association for Information Science and Technology, 65*, 1929–1938.
Wang, D., Song, C., & Barabási, A.-. L. (2013). Quantifying long-term scientific impact. *Science, 342*, 127–132.
Wang, M., Yu, G., An, S., & Yu, D. (2012). Discovery of factors influencing citation impact based on a soft fuzzy rough set model. *Scientometrics, 93*, 635–644.
Wang, M., Yu, G., & Yu, D. (2011). Mining typical features for highly cited papers. *Scientometrics, 87*, 695–706.
Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software, 6*, 3021.
Wooldridge, J., & King, M. B. (2019). Altmetric scores: An early indicator of research impact. *Journal of the Association for Information Science and Technology, 70*, 271–282.
Xu, S., Hao, L., An, X., Yang, G., & Wang, F. (2019). Emerging research topics detection with multiple machine learning models. *Journal of Informetrics, 13*, Article 100983.
Yan, R., Huang, C., Tang, J., Zhang, Y., & Li, X. (2012). To better stand on the shoulder of giants. In *Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries* (pp. 51–60).
Yu, T., Yu, G., Li, P.-Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics, 101*, 1233–1252.
Zhang, L. (2012). Grasping the structure of journal articles: Utilizing the functions of information units. *Journal of the American Society for Information Science and Technology, 63*, 469–480.
Zhao, D., & Strotmann, A. (2020). Deep and narrow impact: Introducing location filtered citation counting. *Scientometrics, 122*, 503–517.